FEBS Letters 584 (2010) 4015–4018

FEBS Letters

journal homepage: www.FEBSLetters.org

Insights into the genomic features and evolutionary impact of the genes configuring duplicated pseudogenes in human

Kamalika Sen, Soumita Podder, Tapash Chandra Ghosh *

Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, India

ARTICLE INFO

Article history:
Received 2 July 2010
Revised 5 August 2010
Accepted 6 August 2010
Available online 12 August 2010

Edited by Takashi Gojobori

Keywords:
Duplicated pseudogenes
Evolutionary rate
Hub-protein
CpG Island
Recombination rate
Functional distance

ABSTRACT

Pseudogenes, regarded as 'genomic fossils', are DNA sequences resembling functional genes in perspective of sequence homology but completely non-functional. In this study, we explored the unique characteristic features of human genes, configuring classical duplicated pseudogenes. We found that progenitors of duplicated pseudogenes are characterized by a high expressivity, and ability to encode hub-proteins in association with a high evolutionary rate. Such unusual features are endorsed by longer protein length, elevated CpG content, and a high recombination rate. The non-functionalization of their duplicated copies can be attributed to the overabundance of gene paralog number in concert with functional redundancy.

1. Introduction

Pseudogenes, the defunct copies of their functional counterparts arise by retrotransposition or duplication followed by various genetic disablements. Due to the shared ancestry with their functional relatives they are considered as 'genetic fossil' and are treated as important resources of comparative genomics [1]. Depending on the origin and characteristic features they are classified as (i) duplicated or non-processed pseudogenes, (ii) processed or retrotransposed pseudogenes, (iii) unitary pseudogenes. The duplicated pseudogenes arise due to unequal crossing over between two homologous chromosomes [2] (during the process of DNA replication) followed by non-deleterious mutations. In spite of having original promoter, intron and exon sequences intact [3], the erroneous recombination and subsequent mutations steer them to the path of non-functionalization. Processed pseudogenes, often termed as "dead on arrival" [4] are ensued by the reverse transcription of mature mRNAs and reinsertion of the cDNAs into the genome [5] whereas unitary pseudogenes are like 'vestigial DNA sequences', which are developed in the genome when a singleton gene is deactivated by mutation that becomes fixed in the population [6]. During the last few decades pseudogenes are being speculated as assets regarding studies of evolutionary relatedness and protein evolution. Pseudogenes, being originated as a consequence of neutral evolution, are often considered as paradigm of neutral evolution [7].

In this study we attempted to draw some probable explanation for the genes targeted to duplicated pseudogenization by their expressivity, hub-protein encoding gene abundance, longer gene length, higher paralog number, abounding CpG content, an elevated recombination rate and functional redundancy. Our study will surely open a new paradigm in duplicated pseudogenization.

2. Materials and methods

Human duplicated gene set and Human pseudogene annotations were retrieved from Ensemble 55 (http://www.ensembl.org/biomart/martview) [8] and pseudogene.org database (Build 36) (http://www.pseudogene.org/) [9], respectively. The number of genes configuring duplicated pseudogenes (i.e. progenitor of duplicated pseudogenes (PDPG)) and the genes casting functional genes (i.e. progenitor of functional genes (PFG)) are 1447 and 2777, respectively (Supplementary Table 1). The corresponding gene sequences were retrieved from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/.

The network connectivity of the genes was obtained from HPRD (Human Protein Reference Database), version 7 (http://www.

Abbreviations: PDPG, progenitor of duplicated pseudogenes; PFG, progenitor of functional genes.

* Corresponding author. Fax: +91 33 2355 3886.
E-mail address: tapash@boseinst.ernet.in (T.C. Ghosh).

0014-5793/$36.00
doi:10.1016/j.febslet.2010.08.012

hprd.org) [10]. Genes having more than five interacting partners were assigned as hub-proteins.

Human duplicated genes and corresponding mouse orthologs were achieved from Ensembl 55. Pair-wise synonymous (dS) and non-synonymous (dN) distances between the orthologous genes of human and mouse were calculated using the PAML package with default parameters [11].

Gene CpG content was estimated by 'cpgreport' (http://www.ebi.ac.uk/Tools/emboss/cpgplot/) [12]. Chromosome wise gene recombination rates were downloaded from http://www.hapmap.org [13]. The recombination rate of the progenitor genes were calculated using the formula $\sum \rho_i/l$, where ρ_i stands for recombination rate at a base position and l for the genic length corresponding to that gene [14].

Gene expression profiles were extracted from Human GeneAtlas GNF1H, MAS5 dataset (http://symatlas.gnf.org) [15]. The average expression values for all probe sets were sorted in ascending order and divided equally into five clusters and were ranked as 1–5, where rank 1 represents lowly expressed genes and rank 5 represents the highly expressed ones.

The functional information carried by the GO annotations was obtained from go molecular functions of human genome in Ensembl 55. The Czekanowski–Dice distance formula used [16] to calculate functional distance of human genes from their corresponding paralogous genes is

$$\begin{aligned}\text{Functional dist}(i,j) = {} & \text{number of } (\text{Terms}(i) \,\Delta\, \text{Terms}(j)) \\ & \setminus [(\text{number of } (\text{Terms}(i) \cup \text{Terms}(j)) \\ & + \text{number of } (\text{Terms}(i) \cap \text{Terms}(j))]\end{aligned}$$

In which, i and j denote two human genes i.e. a gene and its paralogous gene. Terms(i) and Terms(j) are the lists of the GO terms for individual genes. The symbol Δ is the symmetrical difference between the GO term sets of two genes.

3. Results and discussion

3.1. Abundance of hub-proteins and highly expressed genes in the ancestors of duplicated pseudogenes

According to the centrality-lethality rule, highly connected proteins in protein–protein interaction network (hub) are essential for the survival of the organism, since hub-proteins are important for the maintenance of the network structure [17]. In addition, human essential genes are likely to encode hub-proteins and are widely expressed [18]. Investigating the network connectivity of PDPG and PFG, we obtained higher, though not significant, connectivity for PDPG than PFG (average value of interacting partners for PDPG = 9.539 and PFG = 8.86, P value = 1.64×10^{-1} in M-W test). However, we obtained a significantly higher abundance of hub-proteins [Z score = 1.536, confidence level = 93.8%] in the set of PDPG (53.48%) than PFG (47.53%). In addition, highly expressed genes are predominant in PDPG (42.21%) compared to PFG (37.87%) [Z score = 1.325, confidence level = 90.7%]. Thus, the above observations suggest that PDPG act like an essential group of genes.

3.2. Progenitors of duplicated pseudogenes executing high recombination rate harbor a large number of paralogs

It has been suggested that in mammals, protein connectivity is positively correlated with gene duplicability [19]. Thus it is expected that PDPG may exhibit a high paralog number for their hub protein encoding ability. Since, the duplication frequency shares a positive correlation with the recombination rate [20]; PDPG may increase their paralog number by upregulating the recombination rate. Here, we observed a significantly higher recombination rate ($P = 1.37 \times 10^{-183}$ in M-W test, Table 1) for PDPG than PFG. Consequently PDPG exhibit significantly higher paralog number per gene ($P = 2.18 \times 10^{-57}$ in M-W test) (Table 1) compared to PFG and also bear a significant positive correlation between the recombination rate and gene paralog number (Spearman's $\rho = 0.381$, $P = 1.00 \times 10^{-6}$). More interestingly, the paralog number bears a significant but weak positive correlation (Spearman's $\rho = 0.09$, $P = 1.40 \times 10^{-2}$) with the number of duplicated pseudogenes formed per functional gene, which signifies that high paralog number may induce duplicated pseudogenization.

3.3. Evolutionary rate, CpG Island content and length of the genes casting duplicated pseudogenes

It was reported that functionally more important genes will encounter stronger selective pressure than the genes having less functional importance [21]. Research on pseudogenes revealed that, the pseudogenes exhibit extraordinarily high evolutionary rate [22]. So, we were interested to find the evolutionary feature of those genes which are functionally important (higher connectivity, higher expressivity) but give rise to non-functional pseudogenes. Surprisingly, in our analysis we found that PDPG have a significantly higher value of evolutionary rate (dN/dS) ($P = 2.34 \times 10^{-18}$ in M-W test, Table 1) compared to that of PFG. Previously, it was suggested that recombination rate co varies with the rate of neutral mutation [23]. We also noticed a significant positive association between the gene recombination rate and evolutionary rate (Spearman's $\rho = 0.048$, $P = 0.043$) which facilitated to decipher that PDPG, having an elevated rate of gene recombination event, form a number of their duplicated copies which can frequently assemble mutations due to the erroneous duplication process. Another reason for higher substitution rate can be ascribed by the genomic CpG content [24]. Accordingly in our dataset PDPG showed significantly higher values of CpG content ($P = 1 \times 10^{-3}$ in M-W test, Table 1) when compared PFG. Earlier it was reported that gene length can positively constrain the rate of protein evolution [25]. Consistent result has been obtained in our result depicting that the gene length of the highly evolving PDPG is significantly higher ($P = 1.07 \times 10^{-5}$ in M-W test, Table 1) than the slower evolving PFG. Recent discovery on pseudogenes also disclosed that gene length plays an important role in duplicative pseudogenization: longer protein coding genes are more susceptible to produce non-processed pseudogenes as they accumulate more deleterious mutations under a neutral evolutionary scenario [26] which also strengthens our findings.

Table 1
Comparative study between PDPG and PFG.

Average values	Progenitor genes of the duplicated pseudogenes in the paralog set having pseudogenes	Progenitor genes of the paralog set lacking pseudogenes
Gene recombination rate (cM/Mb)	1.442	0.002
Gene paralog number	8.016	2.999
Evolutionary rate (dN/dS)	0.163	0.138
CpG content	136.910	81.195
Gene coding sequence length (bp)	2454.967	1948.845
Functional distance	0.328	0.336

3.4. Functional similarity of the duplicative pseudogene ancestors with other paralog members

The essential nature of PDPG in contrast with their high mutation conceivability induced us to search the exclusive features which can predict the reason behind the non-functionalization of their duplicated gene copies. Recently it has been established that genes forming hub-proteins, sought to retain their homologs providing an extra copy [27] since the researchers argued that genes can maintain redundant copies as 'failsafe' or 'backup' in case the original one is terminated by mutation [28]. Earlier, it was argued that, gene duplication yields functional redundancy and it is often not profitable to retain two identical genes [29] as it enhances metabolic cost for maintaining the genes when accomplishing extra transcription and translation events [30]. To apprehend whether the functional redundancy is responsible for silencing the extra gene copies, we measured the functional distance of the progenitor genes of the non-processed pseudogenes from their paralogous genes. The average value of functional distance of PDPG was found to be significantly lesser ($P = 2.8 \times 10^{-2}$ in M-W test, Table 1) than PFG which indicates that the ancestors of duplicated pseudogenes are functionally closer to their paralogs. The evolutionary rate also exhibits a negative correlation (Spearman's $\rho = -0.024$, $P = 5.9 \times 10^{-5}$) with the functional distance of the above mentioned gene groups indicating the fact that genes with functional redundancy are liable to accumulate non-synonymous base substitutions.

Previously, it was ascertained that the hub-protein encoding genes having higher duplicability [19] execute a higher level of expression [31] which is also consistent with our observations on PDPG. This findings account for a lower evolutionary rate of PDPG. However, in our analysis we observed a higher value of evolutionary rate of PDPG which assigns them as a distinctive gene group with some aberrant characteristics. The selection theory on duplicated genes proposed that, genes, raised by duplication, persist as functional ones, if the mutations create some new genes with some new functions [32]. When the mutations encountered are deleterious then the genes will be lost. But in case the mutations are silent in nature, the duplicated genes will escape the filtration process of natural selection and will be restored in the genome [29]. It is suggested that pseudogenes can serve as a reservoir of sequence variants and can be transferred to the functional genes [33]. Delving deeper into the fact we aimed to picture out the nature of the duplicated genes composing non-processed pseudogenes. We proposed that, the involvement of PDPG in protein–protein interaction network signifies their urge to recombine more frequently in order to increase the paralog number since it is previously suggested that, in mammals, the hub-protein encoding genes display higher gene duplicability by virtue of their need to be produced in a high dosage [19] as well as they may intend to reserve back up copy for future defense. Although the functional similarity of the duplicate genes can provide a back up for gene loss through mutations [34], the redundant copies are not protected against deleterious mutations and thus are evolutionarily unstable [35]. Moreover, PDPG while increasing the paralog number with redundant function may exceed the optimum requirement of the cell. In such a scenario the redundant copies may result into the dosage imbalance which is deleterious for the cellular integrity according to the balance hypothesis [36] of network interacting proteins. Besides that, we here, explained the elevated evolutionary feature of PDPG by their unique genomic features such as longer gene length and affluence of CpG Island. It was also suggested that, in mammals, the enrichment of CpG dinucleotides stimulates gene recombination rate either structurally or by protein binding [37]. In our work the PDPG showing a higher rate of gene recombination and consequently higher duplicability (over PFG) execute an abundance of CpG residues. Our result displaying a positive correlation between gene recombination rate and CpG content (Spearman's $\rho = 0.143$, $P = 1.0 \times 10^{-6}$) also reinforces the aforementioned fact. CpG Island indeed offers the site for mutational hotspot which leads the progenitor genes to gather indels or base substitutions and step up the rate of evolution.

To the best of our knowledge, it is the first detailed structural characterization of the human genes composing non-processed pseudogenes. Our analysis will appreciate the way of future studies on structural and functional characterization of the human genes giving rise to the rest classes of pseudogenes along with their impact on human gene family evolution.

Acknowledgements

Authors are thankful to the Department of Biotechnology, Govt. of India for financial help (sanction number 102/IFD/SAN/PR-1860/2008-09). We thank Mr. S.K. Gupta for his technical helps. We are also thankful to two anonymous referees for their helpful suggestions in improving the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2010.08.012.

References

[1] Zhang, Z.L., Carriero, N. and Gerstein, M. (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. 20, 62–67.
[2] Mighell, A.J., Smith, N.R., Robinson, P.A. and Markham, A.F. (2000) Vertebrate pseudogenes. FEBS Lett. 468, 109–114.
[3] Zhang, Z.L. and Gerstein, M. (2004) Large-scale analysis of pseudogenes in the human genome. Curr. Opin. Genet. Dev. 14, 328–335.
[4] Ding, W.Y., Lin, L., Cheh, B. and Dai, J.W. (2006) L19 elements, processed pseudogenes and retrogenes in mammalian genomes. IUBMB Life 58, 677–685.
[5] Vanin, E.F. (1985) Processed pseudogenes – characteristics and evolution. Annu. Rev. Genet. 19, 253–272.
[6] Zhang, Z.D., Frankish, A., Hunt, T., Harrow, J. and Gerstein, M. (2010) Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol. 11, R26.
[7] Li, W.H., Gojobori, T. and Nei, M. (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292, 237–239.
[8] Hubbard, T.J.P. et al. (2009) Ensembl 2009. Nucleic Acids Res. 37, D690–D697.
[9] Karro, J.E. et al. (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res. 35 (Database issue), D55–D60.
[10] Mishra, G.R. et al. (2006) Human protein reference database – 2006 update. Nucleic Acids Res. 34, D411–D414.
[11] Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.
[12] Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) C_pG islands as gene markers in the human genome. Genomics 13, 1095–1107.
[13] Thorisson, G.A., Smith, A.V., Krishnan, L. and Stein, L.D. (2005) The international HapMap project web site. Genome Res. 15, 1592–1593.
[14] Kato, M., Miya, F., Kanemura, Y., Tanaka, T., Nakamura, Y. and Tsunoda, T. (2008) Recombination rates of genes expressed in human tissues. Hum. Mol. Genet. 17, 577–586.
[15] Su, A.I. et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA 101, 6062–6067.
[16] Baudot, A., Jacq, B. and Brun, C. (2004) A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein–protein interaction network. Genome Biol., 5.
[17] He, X. and Zhang, J. (2006) Why do hubs tend to be essential in protein networks? PLoS Genet. 2 (6), e88.
[18] Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabasi, A.L. (2007) The human disease network. Proc. Natl. Acad. Sci. USA 104, 8685–8690.
[19] Liang, H. and Li, W.H. (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet. 23, 375–378.
[20] Zhang, L.Q., Lu, H.H.S., Chung, W.Y., Yang, J. and Li, W.H. (2005) Patterns of segmental duplication in the human genome. Mol. Biol. Evol. 22, 135–141.
[21] Kimura, M. and Ohta, T. (1974) On some principles governing molecular evolution. Proc. Natl. Acad. Sci. USA 71, 2848–2852.
[22] Miyata, T. and Hayashida, H. (1981) Extraordinarily high evolutionary rate of pseudogenes – evidence for the presence of selective pressure against changes between synonymous codons. Proc. Natl. Acad. Sci. USA 78, 5739–5743.

[23] Hardison, R.C. et al. (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. 13, 13–26.
[24] Subramanian, S. and Kumar, S. (2003) Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 13, 838–844.
[25] Kim, S.-H. and Yi, S.V. (2007) Understanding relationship between sequence and functional evolution in yeast proteins. Genetica 131, 151–156.
[26] Khachane, A.N. and Harrison, P.M. (2009) Strong association between pseudogenization mechanisms and gene sequence length. Biol. Direct 4, 38.
[27] Wu, X.D. and Qi, X.Q. (2010) Genes encoding hub and bottleneck enzymes of the Arabidopsis metabolic network preferentially retain homologs through whole genome duplication. BMC Evol. Biol., 10.
[28] Brookfield, J. (1992) Evolutionary genetics: can genes be truly redundant? Curr. Biol. 2, 553–554.
[29] Zhang, J.Z. (2003) Evolution by gene duplication: an update. Trends Ecol. Evol. 18, 292–298.
[30] Clark, A.G. (1994) Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. USA 91, 2950–2954.
[31] Janga, S.C. and Babu, M.M. (2009) Transcript stability in the protein interaction network of *Escherichia coli*. Mol. BioSyst. 5, 154–162.
[32] Nadeau, J.H. and Sankoff, D. (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics 147, 1259–1266.
[33] Chen, J.M., Cooper, D.N., Chuzhanova, N., Ferec, C. and Patrinos, G.P. (2007) Gene conversion: mechanisms, evolution and human disease. Nat. Rev. Genet. 8, 762–775.
[34] Wagner, A. (2000) Robustness against mutations in genetic networks of yeast. Nat. Genet. 24, 355–361.
[35] Nowak, M.A., Boerlijst, M.C., Cooke, J. and Smith, J.M. (1997) Evolution of genetic redundancy. Nature 388, 167–171.
[36] Papp, B., Pal, C. and Hurst, L.D. (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424, 194–197.
[37] Jensen-Seaman, M.I. et al. (2004) Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. 14, 528–538.